XDOCS: an Application to Index Historical Documents

Federico Bolelli, Guido Borghi, Costantino Grana

Dipartimento di Ingegneria "Enzo Ferrari"
Università degli Studi di Modena e Reggio Emilia
Via Vivarelli 10, Modena MO 41125, Italy
{name.surname}@unimore.it

Abstract. Dematerialization and digitalization of historical documents are key elements for their availability, preservation and diffusion. Unfortunately, the conversion from handwritten to digitalized documents presents several technical challenges.
The XDOCS project is created with the main goal of making available and extending the usability of historical documents for a great variety of audience, like scholars, institutions and libraries. In this paper, the core elements of XDOCS, *i.e.* page dewarping and word spotting technique, are described and two new applications, *i.e.* annotation/indexing and search tool, are presented.

Keywords: Indexing; Page Dewarping; Word Spotting; Word Annotation; Handwriting Recognition.

1 Introduction

The availability of large collection of handwritten historical manuscripts is often required and craved by libraries, scholars and institutions. Despite this, many issues are related to these particular documents.
First of all, the diffusion of historical documents is strictly limited by their physical condition and, often, their are available in a single copy. Moreover, the document readability can be compromised due to the presence of particular handwriting style or other graphic artifacts belonging to old writing techniques.
A solution for these problems can be represented by the dematerialization and digitalization of documents and, in this context, the creation and collection of the so called *Digital Libraries* [5,1,13] represents a key elements in the process of diffusion, usability and availability of historical documents. The XDOCS project is designed with the intention of extending the audience and the access of a huge variety of Italian historical documents.
The conversion from handwritten to digitalized documents represents a great challenge from a technical point of view. On one hand, the peculiarity of this kind of data and the huge amount of documents exclude the possibility to exploit manual annotations and operations which are extremely time-consuming and

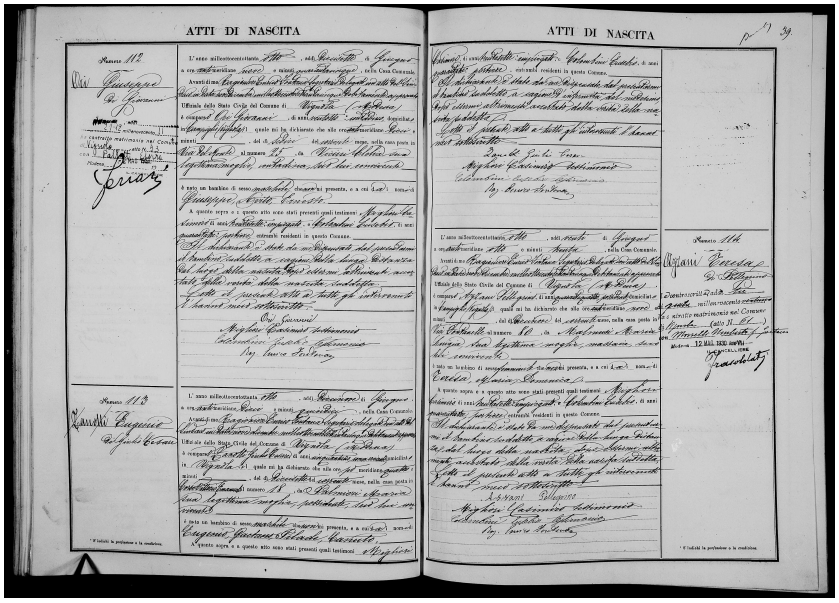


Fig. 1: Example of historical document page dated 1888 and representing three birth acts of the Italian state.

expensive. On the other hand, classical automatic writing recognizers, also called *Optical Character Recognizers* (OCRs), often fail.

One of the first step after the digital acquisition of the document is the page dewarping, since warping distortion affects, as well as the document readability, the performance of automatic techniques of content mining, indexing and word annotation. The input of this process is represented by a curled page (see for instance Figure 1), usually captured by a flatbed scanner and the output is a page containing only horizontal straight lines, without any distortion due to perspective or page warping.

Once historical documents are correctly digitalized and dewarped, it is possible to apply word annotation and word spotting techniques to facilitate the study of researchers and the extraction of semantic contents. A word spotting technique is the ability to create word collections grouped into clusters containing all instances of the same word. Exploiting this technique, it is then possible to index in a semi-automatic way the content of documents.

The paper is organized as follows. Section 2 presents an overall description of related literature works, divided into two main group: page dewarping and word spotting tasks. In Section 3, a general overview of the XDOCS project is given with particular attention to dewarping and word spotting techniques. Section4 describes the annotation/indexing and search tools. All datasets exploited in the project are described in Section 5. Finally, in Section 6 conclusions are drawn.

2 Related Work

Page dewarping. Over the last two decades, many document dewarping techniques have been proposed. The main issue of those proposals is that they are specifically designed for typewritten text, so, they produce low quality results when applied to handwritten texts or hybrid documents (documents that contain a mix of typewritten and handwritten text). Generally, we can divide these approaches in two categories according to the surface model adopted: restoration approaches based on 3D document shape reconstruction [4,7] and restoration approaches based on 2D document image processing [20,8]. The 3D reconstruction models are more accurate but they require images captured with special setup to properly work and this is not the case of common historical digital documents. The 2D approaches, instead, make use of the information contained in the document image in order to restore the page, so they are much more suitable for historical documents. An interesting technique belonging to the second group has been proposed by Stamatopoulos *et al.* in [16]: a two-step approach for efficient dewarping. At the first step, a coarse dewarping is completed with the help of a transformation model, in which a curved surface is projected in a 2D rectangular area. At the second step, fine dewarping is conducted thanks to the word detection, since all words poses are normalized based on the lower and upper word baselines. In [2] a novel approach based on [16] for performing dewarping on Italian historical document images, containing both typewritten and handwritten texts, is presented. This represents the baseline of the XDOCS project so it is described in details in Section 3.1.

Word Spotting. In [11] and [12], the original idea of word spotting for handwritten manuscripts was proposed. In these works the matching techniques and pruning methods are described: given a word image, similar words are clustered and unlikely matches are quickly discarded. Generally, word spotting methods can be divided in two main categories: *line-segmentation* and *word-segmentation* based approaches.

Word-segmentation approaches are based on the hypothesis that each word in the document images is separately clipped. Tomai *et al.* [19] proposed a word-by-word mapping between a scanned document and a manual transcript: in this way, it is possible to perform word location in document pages. This method relies on a *Optical Character Recognizer* (OCR) used as a recognizer for multiple word segmentation hypothesis generated for each line of the document. Results shown that a OCR is not a feasible solution and useful for handwritten historical manuscript recognition. In [15] a local descriptor inspired by a famous key-point descriptor, SIFT [10], is proposed. Here, two different word spotting systems, based on the well-known *Hidden Markov Models* and *Dinamic Time Warping* (DTW), are exploited to achieve significant improvements. In [14] a range of features suitable for DTW has been analyzed. In that paper, different text features, which are extracted from pre-processed rectangular word images and that do not contain ascenders from other words, are used to achieve speed and precision. Exploited features are the gray scale variance, the projection profile, background to ink transitions, the partial projection profile, the upper and lower word pro-

file, and feature sets containing vertical and horizontal partial derivatives. All of them were extracted after normalization of inter-word variations such as *skew* and *slant* angles.

Line-segmentation based methods rely on the hypothesis that each line in the document is separated and word segmentation techniques are not strictly required. Terasawa *et al.* [18,17] presented a word spotting method based on sliding window, line segmentation, continuous dynamic programming and a gradient-distribution-based feature with overlapping normalization and redundant expressions.

In [9] a line-oriented process is applied to avoid the problem of segmenting cursive script into individual words. This approach exploits dynamic programming algorithms and pattern matching techniques. The proposed system is tested on old Spanish manuscripts, showing a high recognition rate. Unfortunately, it is much expensive since words have to be searched for every possible position. Besides, DTW is separately applied on feature vectors and results are merged, producing different alignment for the same word-line pair.

3 The XDOCS Project

Due to the great amount of variability in handwriting styles and the high noise levels in historical documents, handwritten historical documents are generally transcribed by hand. The main goal of the XDOCS project is to develop an innovative data capturing technique able to extract document indexes in quasi-automatic mode from their handwritten contents in order to extend the usability of the historical documents. From a general point of view, the XDOCS application could be split into three main blocks. The first one is the *page dewarping* step during which input digital documents are dewarped and normalized. The second one, the *word spotting* phase, aims at building clusters of words with the same index. Finally, the third block of the project concerns words annotation and smart search of indexes inside the historical documents.

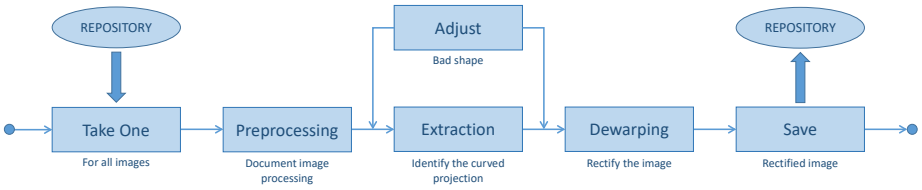


Fig. 2: Pipeline of *Page Dewarping* phase of the XDOCS project.

3.1 Page Dewarping

This step aims to transform the original curled document pages into ones constituted only of horizontal straight text lines, without any distortion caused by

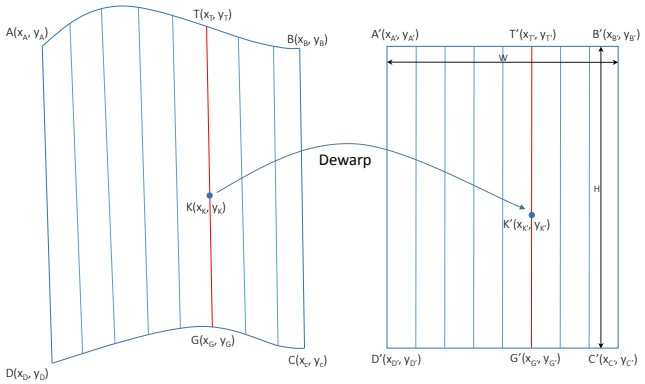


Fig. 3: Dewarping transformation model: projection of the curved surface on the left side, 2D rectangular destination area on the right side.

lenses and perspective. *Page dewarping*, depicted in Figure 2, is essentially composed by three steps:

- *Image pre-processing* step consists in filtering out document and page noises, mainly caused by the intrinsic nature of the original images which belongs to old documents and the digitization process. This filtering relies on connected components statistics as described in [2].
- *Projection extraction* module aims to identify the curved 2D projection surface surrounding the document page. According to the warping model, the projection is assumed to be described by two almost vertical straight lines ($y = ax+b$) and by two third degree polynomial curves ($y = ax^3+bx^2+cx+d$). The vertical lines are automatically identified by the use of the *Hough* transform [6], while the coefficients of polynomial lines are fitted with the Least Square Estimation algorithm. Boundary extraction significantly influences the quality of the dewarping process, and then the entire pipeline of the XDOCS application: if it fails the *Adjust* step leaves the user the possibility to correct curves via a GUI.
- *Dewarping* is the core of the image rectification process. During this phase, the projection of the curved surface is mapped into a rectangular normalized 2D area. The transformation is described by Equation 1 where, referring to Figure 3, $|AD|$ and $|BC|$ are euclidean distances and $|\widehat{AB}|$ and $|\widehat{CD}|$ are the length of polynomial curves on the projection surface. Moreover, the two points T and G belong respectively to the curves $|\widehat{AB}|$ and $|\widehat{DC}|$. The idea is to preserve proportions between dimensions of projected curves and 2D destination area.

$$K'(x'_A + W * \frac{|\widehat{AT}|}{|\widehat{AB}|}, y'_A + H * \frac{|TK|}{|TG|}) \tag{1}$$

3.2 Word Spotting

At the end of *Page Dewarping*, input images are correctly rectified: they do not suffer of any distortion effects, they are normalized to fixed dimensions and are then ready for the *Word Spotting* [3].

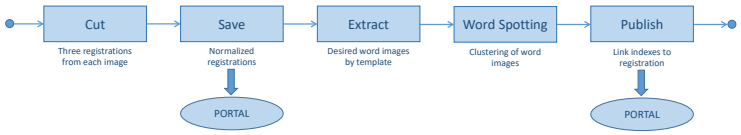


Fig. 4: Pipeline of *Word Spotting* phase of the XDOCS project.

The aim of this step is to group all words of interest into clusters in order to greatly reduce the amount of annotation work that has to be performed in the last phase of the XDOCS project. *Word Spotting* articulates as follows:

- *Cut* three registrations from each image in order to logical separate document contents. This step will produce a series of images like the one reported in Figure 5. Each resulting image will be stored in a database for user consultation.
- Words representing intended indexes are then *Extracted* exploiting a simple template approach: given that all acts have the same structure (for a give historical book) and were normalized in the previous step, the extraction template can be defined once for all documents.
- *Word Spotting* is the core step of the current process. Firstly, word images are preprocessed and normalized as described in [3]. Then, HOG feature vectors are extracted from each word image exploiting a sliding window approach. Finally, words are matched and grouped together using the similarity distance obtained with DTW technique.

3.3 Indexing and Search Tools

The third part of the XDOCS project is constituted by the annotation/indexing tool and by the advanced search system. Since they are the main novelty of this work they are described in detail in Section 4.

4 XDOCS Application

4.1 Annotation/Indexing Tool

The annotation system is based on the word spotting approach described in the Section 3.2. When a new registry is loaded onto the system it is processed following the pipeline of Figure 4 and all extracted words associated to the same

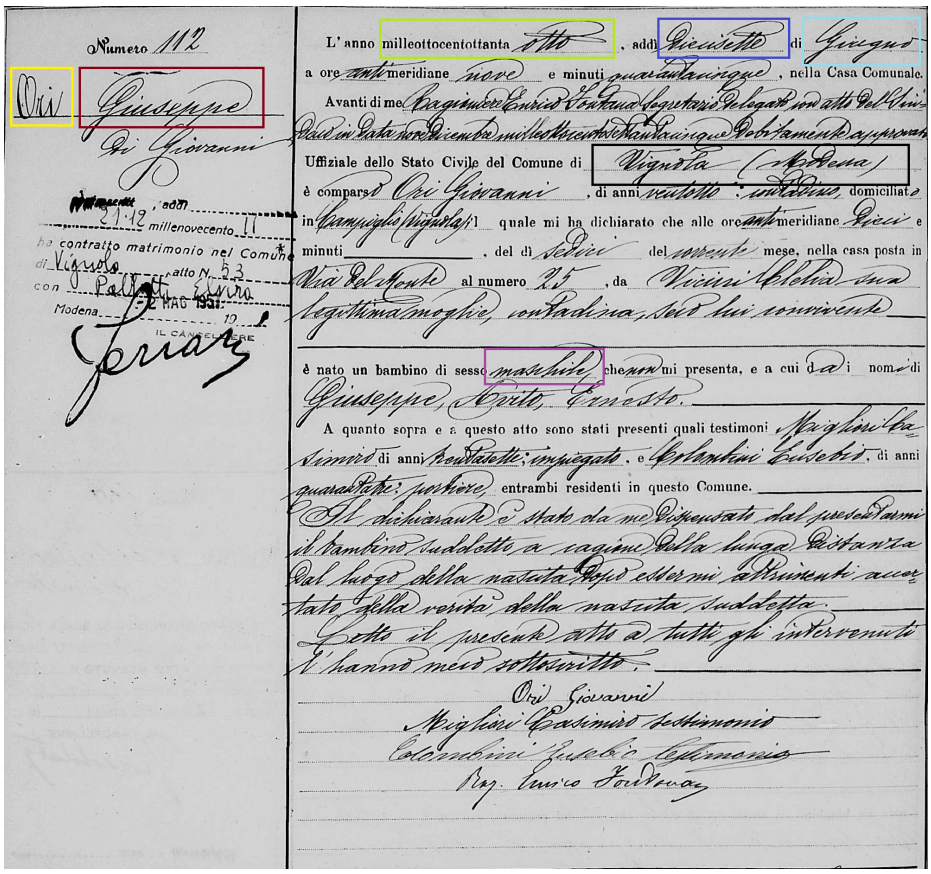


Fig. 5: Example of single birth act registration obtained after the *Cut* step of *Word Spotting*. The colored bounding boxes contain searched indexes and are automatically extracted by a template approach.

index are compared together. After that, it is possible to build a ranking system and store into the database all matches between a word and those most similar to it. The association list can be consulted by the users, as shown in Figure 6. The tool provides, for each word/index the following information:

- General information about the document from which the word belongs to;
- The word image of the index and the associated plain text value, when available;
- A list of most similar words: for each of them, the distance calculated by the word spotting algorithm, the associated image, and a check box to specify if the indexes are actually equivalent are reported.

Thanks to the annotation interface it is possible to set or update information automatically extracted by the *Word Spotting* algorithm and propagate them

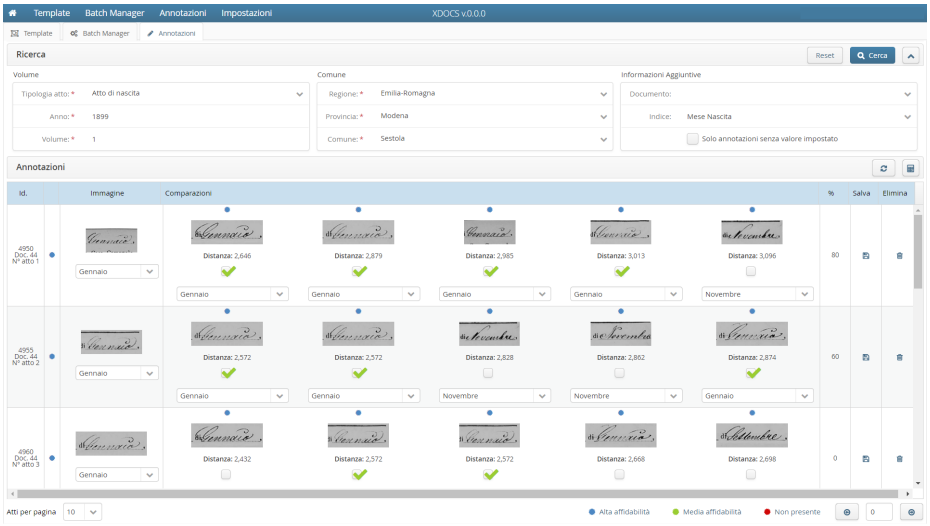


Fig. 6: Annotation interface of the XDOCS project

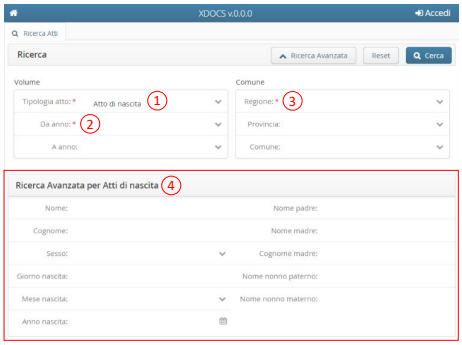
to the associated words. This procedure allows to significantly reduce time and costs of the annotation process improving the quality and the performance of the search system described in the following section.

4.2 Search Tool

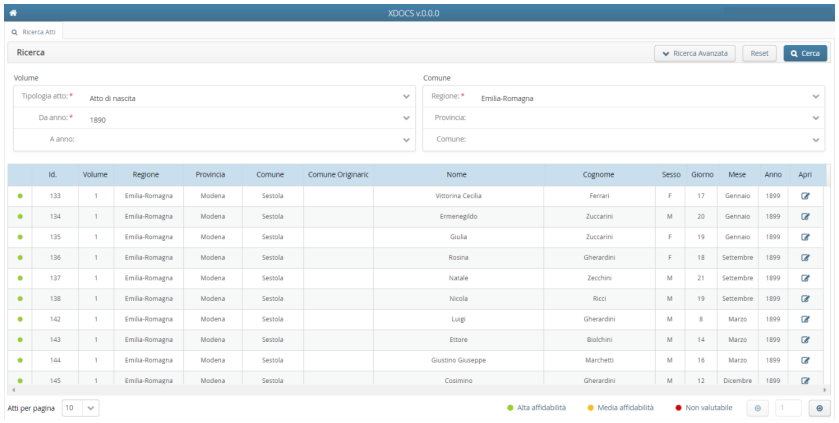
In order to achieve the goal of simplifying accesses to historical documents, a browser interface based on *PostgreSQL* database has been developed. The software provides an advanced search tool which allows the user to search single act page specifying, for example, the type (*1*), the year (*2*) and the municipality of the searched act (*3*) (see for instance Figure 7a). Additional search fields (*4*) are available after selecting the act type. Figure 7a reports an example of advanced search fields specific for the birth act: name, surname, sex, day/month/year of birth, father name, mother name and so on. It is important to highlight that all search parameters, but the act type, the year and the municipality, are optional, and the *full text search* of *PostgreSQL* backend will combine them to provider the best search results.

The acts identified by the search process will be displayed in list, ordered by id (Figure 7b). As depicted in see Figure 7c, it is possible to select and display one specific act and all the associated words which are classified in three categories distinguished by colors:

- Red: words which require annotation;
- Green: words annotated by user without administrator privileges;
- Blue: words specified by an administrator user.



(a) Advanced search page.



(b) Search results page.

(c) Indexes associated to a birth act.

Fig. 7: XDOCS search tool.

Fig. 8: Inter dataset variations, *i.e.* the same month name written by different writers.

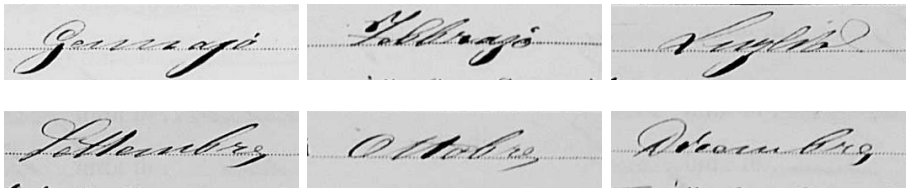


Fig. 9: Intra dataset variations, *i.e.* different months written by the same writer.

Additionally, through this interface, it is possible to set and change the value of indexes: any change can be automatically propagated depending on user permissions.

5 XDOCS Dataset

As mentioned above, XDOCS is designed with the intention of extending to a much wider audience the possibility to access a variety of historical documents. To that purpose, a great amount of Italian historical birth certificates documents of the XIX century has been collected.
Moreover, to test and evaluate the proposed tools, *Word Spotting* and *Annotation*, a huge collection of single word images has been collected and annotated. All these datasets are publicly released [1].

5.1 Personal Information Data

This is a newly collected dataset consisting of annotated words images of names, surnames, birthdays, municipalities and sex (see Figure 10 for instance). All images are taken from Italian civil registries of the XIX century. Writing styles variety is guaranteed due to the presence of different writers.

5.2 Months Data

This sub-dataset is firstly introduced in [3] and consists of a collection of handwritten month names extracted from Italian civil registries of the XIX century. Specifically, the dataset counts 1200 words of all 12 months: data are affected by both *intra* and *inter* variations, due to the presence of three different official state writers, as depicted in Figure 8 and 9.

[1] http://imagelab.ing.unimore.it/XDOCS

Fig. 10: Examples of intended indexes extract using a template approach from normalized registries. In turn they are: day and year of birth, municipality, name, surname and sex.

6 Conclusions

In this paper we describe the XDOCS project, that includes the page dewarping and the word spotting techniques. Moreover, two frameworks are introduced and described. The first one, the *Annotation* tool, is created to facilitate the annotation of words belonging to historical handwritten documents and the second one, the *Search* tool, is designed to allow searching of these words.
The XDOCS project has the main goal of encouraging the diffusion of handwritten historical documents, generally characterized by difficulties in readability, comprehension and physical availability.

Acknowledgement

The XDOCS project is currently underway at SATA s.r.l. in collaboration with the University of Modena and Reggio-Emilia, and co-funded by the Emilia-Romagna regional administration.

References

1. Balducci, F., Borghi, G.: An annotation tool for a digital library system of epidermal data. In: Italian Research Conference on Digital Libraries. pp. 173–186. Springer (2017)
2. Bolelli, F.: Indexing of historical document images: Ad hoc dewarping technique for handwritten text. In: 13th Italian Research Conference on Digital Libraries. IRCDL (February 2017)
3. Bolelli, F., Borghi, G., Grana, C.: Historical handwritten text images word spotting through sliding window hog features. In: 19th International Conference on Image Analysis and Processing (2017)
4. Cao, H., Ding, X., Liu, C.: Rectifying the bound document image captured by the camera: A model based approach. In: Document Analysis and Recognition, 2003. Proceedings. Seventh International Conference on. pp. 71–74. IEEE (2003)

5. Corbelli, A., Baraldi, L., Balducci, F., Grana, C., Cucchiara, R.: Layout analysis and content classification in digitized books. In: Italian Research Conference on Digital Libraries. pp. 153–165. Springer (2016)
6. Duda, R.O., Hart, P.E.: Use of the hough transformation to detect lines and curves in pictures. Communications of the ACM 15(1), 11–15 (1972)
7. Fu, B., Wu, M., Li, R., Li, W., Xu, Z., Yang, C.: A model-based book dewarping method using text line detection. In: Proc. 2nd Int. Workshop on Camera Based Document Analysis and Recognition, Curitiba, Barazil. pp. 63–70 (2007)
8. Gatos, B., Pratikakis, I., Ntirogiannis, K.: Segmentation based recovery of arbitrarily warped document images. In: Ninth International Conference on Document Analysis and Recognition (ICDAR 2007). vol. 2, pp. 989–993. IEEE (2007)
9. Kolcz, A., Alspector, J., Augusteijn, M., Carlson, R., Popescu, G.V.: A line-oriented approach to word spotting in handwritten documents. Pattern Analysis & Applications 3(2), 153–168 (2000)
10. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International journal of computer vision 60(2), 91–110 (2004)
11. Manmatha, R., Croft, W.: Word spotting: Indexing handwritten archives. Intelligent Multimedia Information Retrieval Collection pp. 43–64 (1997)
12. Manmatha, R., Han, C., Riseman, E.M., Croft, W.B.: Indexing handwriting using word matching. In: Proceedings of the first ACM international conference on Digital libraries. pp. 151–159. ACM (1996)
13. Pini, S., Cornia, M., Baraldi, L., Cucchiara, R.: Towards video captioning with naming: A novel dataset and a multi-modal approach. In: International Conference on Image Analysis and Processing. pp. 384–395. Springer (2017)
14. Rath, T.M., Manmatha, R.: Features for word spotting in historical manuscripts. In: Document Analysis and Recognition, 2003. Proceedings. Seventh International Conference on. pp. 218–222. IEEE (2003)
15. Rodriguez, J.A., Perronnin, F.: Local gradient histogram features for word spotting in unconstrained handwritten documents. Proc. 1st ICFHR pp. 7–12 (2008)
16. Stamatopoulos, N., Gatos, B., Pratikakis, I., Perantonis, S.J.: A two-step dewarping of camera document images. In: Document Analysis Systems, 2008. DAS'08. The Eighth IAPR International Workshop on. pp. 209–216. IEEE (2008)
17. Terasawa, K., Nagasaki, T., Kawashima, T.: Eigenspace method for text retrieval in historical document images. In: Document Analysis and Recognition, 2005. Proceedings. Eighth International Conference on. pp. 437–441. IEEE (2005)
18. Terasawa, K., Tanaka, Y.: Slit style hog feature for document image word spotting. In: Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on. pp. 116–120. IEEE (2009)
19. Tomai, C.I., Zhang, B., Govindaraju, V.: Transcript mapping for historic handwritten document images. In: Frontiers in Handwriting Recognition, 2002. Proceedings. Eighth International Workshop on. pp. 413–418. IEEE (2002)
20. Ulges, A., Lampert, C.H., Breuel, T.M.: Document image dewarping using robust estimation of curled text lines. In: Eighth International Conference on Document Analysis and Recognition (ICDAR'05). pp. 1001–1005. IEEE (2005)